WEB SCRAPING WITH KNIME

Introduction:-

Web scraping is the most important thing in Data science. our objective in this article is to make a simple workflow in knime scraps the data from the website. Here we have to take a live website of Corona cases and other information for India, you can find the website here.

X path:-

* So our next step is to parse the XML file to extract the table data from it. X-path node helps you out to parse the XML file, as you can see below the configuration of xpath we have to specify the X path

* Using the Add path button you can add a different path and X path summary displays how many paths you have.

pivoting and ungroup:-

Now our next step is to transform the table into a data file we use pivoting node. And After this, we have to ungroup the column using ungroup node.

Webpage Retriever Node:-

* Web page Retriever Node provides the facility to Connect with the website and generate the XML file, so as you can see in the below dialog box of webretriever node you have to just specify your respective URL in the connection setting in our case we have to specify this web site in URL

* You can find a more detailed description of webpage Retriever Node from here.

CSV Writer:-

our final step is to write a excel (or) CSV writer file to this fetched data and Knime provides facilities to do this by simply using CSV writer nodes.

So by executing this full Workflow at the end in the CSV or excel you get the fetch data a data - wise bitcoin rest.

CORRELATION

Correlation Analysis

Correlation is a statistical technique that can be used to determine if, how strongly pairs of variables are associated. Correlation is only appropriate for quantifiable data in which numbers are meaningful, such as continuous or ordinal data. It cannot be used for purely categorical data for which we have to use contingency table analysis. If one variable deviates from its mean does the other variable deviate from its mean in either the same or opposite direction. This can be assembled by measuring covariance, however, this not standardised. We can measure the covariance of two variables which are measured in meters, if we convert the same values to centimetres, we get the same relationship but which a completely different covariance value. In order to overcome this, standardised covariance is used which is known as pearsons correlation coefficient (or "r") it ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. If r is

Close to 0, there is no relationship. If r is (+) then as one variable increases the other also increases. If r is (-) then as one increase. If r is (-) then as one increases, the other decreases (sometimes referred to as an "inverse correlation)

The correlation coefficient (r) should not be confused with R^2 (coefficient of determination) or R (multiple correlation coefficient as used in regression).

The main assumption in this analysis is that the data have a normal distribution and are linear. This analysis will not work well with curvilinear relationships.

Let's run the correlation analysis to investigate whether the correlation coefficient between the covert and overt antisocial behaviour (convert and overt) is significantly different from 0.

From now on, we will use the alpha level of .05 because it is most commonly used.